Néologie, classes d'objet et extraction automatique

Jean-François Sablayrolles

HAL JFS 34 2010a

Sablayrolles Jean-François (Paris 13 et LDI UMR 7187)

« Néologie, classes d'objets et extraction automatique »,

Actes del I Congrés Internacional de neologia de les Llengües Romàniques (Barcelone (7-10 mai 2008), M. Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa, Mercè Lorente eds., Barcelone, Université Pompeu Fabra, IULA, sèrie activitats 22, p. 143-149

Résumé

Les extracteurs automatiques de néologismes pêchent souvent par excès ou par défaut : par excès car les corpus d'exclusion sont incomplets, par défaut car sont omis un grand nombre de néologismes (néologismes polylexicaux, néologismes par lexicalisation de séquences d'origine syntaxique, néologismes sémantico-syntaxiques, néologismes formels homonymiques de mots existants). Un certain nombre d'aménagements permettent de réduire les inconvénients mais c'est sans doute un modèle linguistique tel celui des classes d'objets qui fournit les bases les plus fermes pour asseoir une extraction automatique de tous les néologismes (avec la notion d'emploi non décrit dans les dictionnaires électroniques). En attendant que de tels dictionnaires soient achevés, sont présentés les outils mis en œuvre dans l'équipe néologie du LDI.

Mots-clés : néologie, extraction automatique, classes d'objets, veille néologique, types de néologismes

NEOLOGISMES, CLASSES D'OBJETS ET EXTRACTION AUTOMATIQUE

Introduction

L'extraction des néologismes et la conception de la néologie[1] et des critères de néologicité sont interdépendants. La conception de la néologie influence les incorporations qui, en retour, conduisent à affiner le concept. S'instaure ainsi un mouvement dialectique entre théorie et pratique. Les outils que l'on a à disposition ou que l'on se forge ont également des incidences sur le travail relatif aux néologismes et le développement des ressources informatiques modifie sensiblement les modes de la veille néologique. Certaines lacunes de l'extraction automatique peuvent être comblées par le recours à des dictionnaires formés sur le modèle des classes d'objets. Les outils et les objectifs de l'équipe néologie du LDI jouent également un rôle tant dans la conception de la néologicité que dans les incorporations.

1. Apports et faiblesses de l'extraction automatique

L'accès à de grand corpus numérisés, en particulier de presse, ainsi que l'existence de dictionnaires informatisés permettent de traiter, sans risque d'étourderie, des données incomparablement plus abondantes qu'auparavant et de livrer des candidats néologismes en grand nombre — et même en trop grand nombre vu l'importance numérique des fautes de frappe, des fautes d'orthographe et autres problèmes matériels, sans compter les noms propres de tous types. Des filtres, comme ceux construits naguère (Mathieu *et alii*, 1998) ou actuellement au laboratoire, doivent être élaborés pour diminuer le bruit.

[1] Terme employé ici dans l'acception traditionnelle d'ensemble des procédés de formation de nouvelles unités lexicales plus que dans les acceptions plus récentes évoquées, lors de la séance inaugurale, par Bernard Quemada qui constate et regrette leur oubli dans les dictionnaires contemporains.

Mais à côté du trop plein, il y a aussi les lacunes de l'extraction automatique fondée sur le seul critère d'exclusion lexicographique. Elles sont bien connues : outre le problème difficile du choix du/des dictionnaires de référence[2], on ne collecte que des néologismes formels (et encore pas tous) : manquent :

a) des néologismes polylexicaux (sans traits d'union) : *embrasement généralisé éclair* (*flash over*)[3], *lanceur d'alerte*[4]…

b) des néologismes purement syntaxiques :

b1) par changement de construction : *flasher* connaît conventionnellement des constructions transitives directes (*flasher un texte*, ou *une voiture en excès de vitesse*) ou indirectes (*flasher sur quelqu'un ou quelque chose* 'avoir le coup de foudre pour'), mais pas de constructions intransitives comme : *La vitrine de cette boutique flashe d'emblée* 'attire le regard'[5] ;

b2) par conversion : *ça m'esclave sévère* « cela me rend fortement esclave => je suis très dépendant de » où *esclave*, originellement nom ou adjectif, fonctionne comme un verbe et où l'adjectif *sévère* fonctionne comme un adverbe[6] ;

c) des néologismes sémantico-syntaxiques :

c1) par extension d'emploi : *incuber une entreprise* (« Quinze entreprises ont été incubées », comme des bébés prématurés pour les aider à vivre et se développer[7]) ;

c2) par un emploi imagé : *Bounty* 'immigré de la 2e génération bien intégré' car noir dehors et blanc dedans (comme la confiserie), qualificatif donné par des élèves à leurs professeurs issus des mêmes quartiers défavorisés qu'eux[8] ;

d) des néologismes formels homonymes de mots conventionnels :

d1) avec une analyse morphologique différente : *rosac-é* 'en forme de rosace' et *ros-acé* 'd'une sorte de couleur rose'[9]. *Carté* 'qui a une carte d'abonnement de cinéma' est nouveau et différent du participe du verbe (rare) *carter* 'présenter quelque chose sur une carte' : *des boutons cartés*[10] ;

d2) avec un changement de valeur dû à une ellipse (le résultat se présentant comme une conversion) : *portable* (par ellipse de *téléphone*), *argentique* (par ellipse de *appareil photo*) ;

d3) par fabrication sur un homonyme ou sur une autre acception d'un mot polysémique : *frictionner* 'frotter, faire une friction' est conventionnel mais *ça*

---

[2] Que faire par exemple pour des unités absentes du *PLI* et du *PR* mais qui ont une entrée en tant que telle dans Wikipédia (avec une définition, une histoire du mot, un équivalent dans une langue étrangère, l'anglais le plus souvent, etc.) et dont les moteurs de recherche donnent un nombre non négligeable d'occurrences dans les pages françaises, comme c'est le cas pour *lanceur d'alerte* par exemple ?

[3] *Direct soir*, 14/02/2008.

[4] *Métro*, 14/03/2008.

[5] Je remercie Jean-Claude Boulanger qui m'a signalé que cette construction est attestée de longue date en français québécois. La nouveauté doit donc s'apprécier par rapport à des usages en vigueur à tel ou tel endroit, à telle ou telle époque. Et, malgré leurs lacunes ou imperfections, les dictionnaires sont des outils incontournables et en fait assez fiables pour connaître ces usages. On peut en effet penser que l'absence de l'indication de constructions intransitives tant dans le *PLI* que le *PR* correspond à une absence dans l'usage régulier du français hexagonal.

[6] Exemple déjà un peu ancien (milieu des années 90) d'un slogan publicitaire pour des yaourts, sur une affiche de Claire Brétécher.

[7] *20 minutes*, 31/01/2007.

[8] *20 minutes*, 14/12/2007. Cette antonomase est fondée sémantiquement sur une métaphore.

[9] Lu, il y a longtemps, dans une copie d'étudiant. Le fait que ce soit une « faute » n'ôte en rien le caractère nouveau de cet emploi.

[10] *Télérama*, 05/09/2007.

frictionne ‘il y a des frictions, des tensions’ (dans une association) est nouveau[11] ;

e) des néologismes avec un sens compositionnel différent du sens conventionnel : *intraitable* ‘qui ne peut pas être traité’ et non ‘insensible, impitoyable’[12].

2. Des améliorations en cours

2.1. Des analyseurs morphologiques et syntaxiques

Des analyseurs morphologiques et syntaxiques permettent de relever certaines de ces innovations — en particulier les conversions, puisque les emplois nouveaux ne peuvent pas coïncider avec les parties du discours auxquelles appartiennent traditionnellement ces unités lexicales — mais pas toutes. La néologie sémantico-syntactique en particulier résiste.

2.2. La théorie des classes d’objets

La théorie des classes d’objets développée par Gaston Gross (1994 entre autres) ouvre une voie à une définition plus satisfaisante de la néologie, et, à plus long terme, à des possibilités d’extraction automatique. Une thèse, dirigée par Salah Mejri, est en cours au laboratoire LDI dans ce domaine. C’est moins en effet le concept de néologie qui pose un problème[13] que l’insuffisance de descriptions complètes des langues, et surtout, pour l’extraction automatique des néologismes, de descriptions entièrement formalisées. Est en effet néologisme toute création formelle (nouveau signifiant ou homonyme) ou toute innovation dans l’utilisation d’une unité lexicale[14] par rapport au savoir des locuteurs natifs tel qu’il doit être formalisé dans des dictionnaires électroniques. Ceux-ci se présentent comme des dictionnaires de phrases élémentaires, c’est-à-dire de prédicats saturés par leurs arguments. Prenons l’exemple du prédicat *prendre* sous sa forme verbale ou sous sa forme nominale *prise*[15].

prendre N0 <hum>, N1<lieu, ville> *Troie*
prédicat nominal *la prise de Troie*
prendre N0 <hum>, N1<voie de communication> *la rue*
*prédicat nominal **la prise de la rue*
prendre N0 <hum>, N1<moyen de transport> *la voiture*
*prédicat nominal **la prise de la voiture*
prendre N0 <hum>, N1 <boisson> *une bière*
*prédicat nominal **la prise d’une bière*
prendre N0 <hum> N1 <médicament> *un cachet*
prédicat nominal *la prise d’un cachet*

Un slogan publicitaire pour la SNCF[16] viole délibérément ce schéma en étendant au prédicat nominal une construction possible uniquement avec le prédicat verbal dans

[11] *Le réveil cantalien*, 23/11/2007.
[12] Cet exemple est un peu ancien. Ce phénomène de la possibilité toujours ouverte de donner un sens compositionnel à des lexies complexes démotivées a été bien décrit par Danielle Corbin (1987, 1988 entre autres) qui en a fait un argument en faveur du modèle de morphologie associative et stratifiée qu’elle a développé.
[13] Ce qui pose un vrai problème, comme l’a montré Jean-Claude Boulanger dans la conférence de clôture, c’est la durée du sentiment néologique à propos d’un néologisme. Elle est due à plusieurs facteurs et n’est pas décidable automatiquement et uniformément.
[14] Nous entendons par là, à la suite de la définition de la lexie par Bernard Pottier, un signe linguistique qui est une unité fonctionnelle et qui est mémorisable en compétence.
[15] Pour une présentation plus détaillée de ce modèle — et, en particulier, pour le fait que les prédicats ne sont pas nécessairement verbaux et que tous les verbes ne sont pas des prédicats —, nous renvoyons à la conférence plénière de Salah Mejri lors de ce congrès.
[16] Dans *Le Monde*, 23/06/2006.

La prise de train bénéficie à la santé de votre voiture

Cette transgression ludique, qui relève de la néologie sémantico-syntaxique avec un emploi non conventionnel et innovant, ne peut être repérée qu'avec un dictionnaire qui donne la description fine et complète de tous les emplois d'une unité lexicale. Il en va de même avec

Récolter le vent[17]

car *vent* appartient à la classe <événement atmosphérique> qui ne fait pas partie des schémas argumentaux conventionnels de *récolter* qui pourraient être schématiquement décrits ainsi :

récolter N0<hum>, N1inc <produit de la terre> 'recueillir' *du blé, des navets*…
N1évé <ennuis> 'obtenir' *des désagréments, des problèmes*…
N1évé <coups> 'recevoir' *une gifle* (familier)

L'aspect frappant du titre[18] récent, *Marc Machin encourt la liberté*, est fondé également sur une rupture par rapport aux classes d'objets que peut avoir le prédicat verbal *encourir* comme deuxième argument : uniquement des événements fâcheux, en particulier des peines prononcées par des tribunaux. Or la *liberté* n'entre pas dans ces classes.

Le concept de classe d'objets permet donc de fonder théoriquement le concept de néologie et les dictionnaires fondés sur ce principe peuvent servir efficacement de corpus d'exclusion.

3. Objectifs et outils de l'équipe néologie de LDI

3.1. Le partage des tâches

Mais il ne faut pas se cacher que ces outils ne seront pleinement opérationnels que le jour où l'ensemble de la langue aura été décrite ainsi d'une manière complètement formalisée. En attendant, des dépouillements manuels restent incontournables, mais ils sont allégés de la tâche du relevé des néologismes monolexicaux formels, sauf d'éventuels homonymes, et se concentrent sur les autres formes de néologie.

3.2. Deux outils informatiques : un extracteur et une base de données

Au laboratoire LDI, l'équipe néologie, et deux de ses informaticiens ont développé récemment, en étroite collaboration, deux outils (présentés lors de ce congrès) : Néologia, une base de stockage et de traitement de néologismes, et Télanaute qui arpente la toile et livre des candidats néologismes avec le corpus d'exclusion Morfetik qui compte environ un million de formes fléchies. Quatre voies s'ouvrent à ces candidats :

i) inclusion dans le dictionnaire Morfetik en cas de lacune de celui-ci : *glamour* ;

ii) mise dans des filtres divers (toponymes, anthroponymes, fautes récurrentes comme *infractus* et toutes les fautes d'accent…) ;

iii) sas ou dictionnaire annexe pour des lexies qui circulent trop pour être encore néologiques et donc être entrées dans la base Neologia mais qui n'ont pas encore non plus une diffusion générale qui les fasse intégrer dans le dictionnaire Morfetik : *tektonik* (type de danse), une grande partie du vocabulaire du clubbing, des jeux électroniques (*CLUF* 'contrat licence utilisateur final'), etc. ;

iv) entrée dans la base de données néologiques Neologia, qui comporte une vingtaine de champs répartis en 5 grands groupes, les 4 premiers liés à l'item et le 5^{e} au contexte.

---

[17] *20 minutes*, 05/02/2007, gros titre en première page sur fond d'éoliennes transformant le vent en énergie électrique.

[18] *20 minutes*, 05/05/2008. Il s'agit d'un prisonnier condamné pour un meurtre dont s'accuse depuis quelqu'un d'autre. La révision de son procès pourrait l'innocenter et le faire libérer.

Une présentation plus détaillée de ces deux outils étant effectuée ailleurs dans les actes du congrès (Fabrice Issac et Soundous Ben-Hariz Ouenniche pour Télanaute et Emmanuel Cartier et Jean-François Sablayrolles pour Neologia) nous passons au dernier point, à savoir que l'extraction ne peut pas se faire sans la définition d'un projet. Trois types de choix ont été opérés par notre équipe.

3.3. Veille et analyse

L'organisation de la base Neologia est liée au fait que l'objectif visé n'est pas la seule collecte de néologismes mais aussi leur analyse, à l'aide de requêtes simples et de requêtes croisées mettant en jeu plusieurs champs simultanément comme « partie du discours », « matrice lexicale », « type d'émetteur », « domaine du savoir », etc. L'objectif est de mieux savoir quel type d'émetteur crée quel type de néologisme et dans quel type de situations. Est-il besoin de dire que ce type d'études est grandement facilité par le recours à des bases de données informatisées ?

3.4. Deux types de corpus

Par ailleurs, la veille néologique peut avoir deux objectifs principaux : l'étude de l'évolution du lexique d'une langue à un moment donné de son histoire, auquel cas on ne relève que des lexies qui connaissent une circulation significative, celles qui sont sur le point (ou qui viennent) d'entrer dans des dictionnaires généraux. Mais on peut aussi viser l'étude de la créativité lexicale des membres d'une communauté linguistique. Dans ce cas, on relève toutes les créations, même si elles demeurent des hapax ou sont de diffusion limitée[19].

3.5. Langue générale et domaines de spécialité : des frontières non étanches

L'équipe s'intéresse par ailleurs à la langue générale à l'exclusion des vocabulaires de spécialité qui ont un fonctionnement un peu différent, mais, là encore, les choses ne sont pas simples pour l'extraction automatique sur le seul critère d'exclusion lexicographique puisqu'il y a des échanges dans les deux sens[20]. Si nos corpus de base relèvent de la langue générale, avec en particulier certains titres de presse généralistes, il reste que les termes ne sont pas complètement exclus de nos préoccupations. On relève par exemple des passages d'un domaine de spécialité à la langue générale. Il s'agit du phénomène de la divulgation (*vraquier* 'cargo qui transporte de la marchandise en vrac' ne s'est répandu qu'après le naufrage d'un navire de ce type) ou d'emploi imagé comme *tacler* un adversaire politique (et non un joueur de foot). On relève également les passages d'un domaine spécialisé connu à un autre, comme d'un sport à l'autre. En revanche le passage de mots de la langue générale à un domaine de spécialité pointu ne nous retient pas. Il s'agit d'un type particulier de néonymie, comme *chaussette* dans le domaine du nucléaire par exemple.

Conclusion

Reste, comme l'a rappelé Pierre Auger lors de la table ronde, le problème de la variation du sentiment néologique d'un collecteur à un autre (et aussi chez un même collecteur au cours du temps). Des membres de notre équipe se sont livrés à une expérimentation de collecte et d'analyse en quadruple aveugle afin d'identifier les causes des fluctuations pour les réduire le plus possible et homogénéiser les résultats, condition nécessaire à toute étude ultérieure significative sur la néologie.

---

[19] Mais on n'est jamais sûr de l'avenir d'une lexie, et certaines peuvent connaître des diffusions tardives.

[20] Louis-Jean Rousseau a proposé lors d'une table ronde que les corpus d'exclusion soient élargis et prennent en compte aussi les dictionnaires terminologiques. De fait un certain nombre de « néologismes » existent déjà dans des domaines de spécialité et ne sont pas alors radicalement nouveaux. Mais la pénétration de termes dans la langue générale les rend néologiques pour les usagers de celle-ci quand ils les découvrent dans un média généraliste.

Les discussions et le travail définitoire ont permis de rapprocher les points de vue et deux des quatre collecteurs parviennent à un taux d'accord supérieur à 95%, ce qui est plutôt satisfaisant[21].

Les multiples références de cet article à d'autres interventions du congrès montrent à quel point les rencontres et les discussions permettent d'enrichir les points de vue et combien ce congrès CINEO, le premier congrès international de la néologie des langues romanes, a été fructueux. Que Térésa Cabré et son équipe organisatrice en soient chaleureusement remerciées

Indications bibliographiques

BENHARIZ OUENNICHE Soundous, , « Vers une homogénéisation des incorporations des néologismes », *Neologica* 3, Classiques Garnier, 2009.

CORBIN Danielle, 1987, *Morphologie dérivationnelle et structuration du lexique*, 2 vol., Tübingen, Max Niemeyer Verlag.

CORBIN Danielle, 1988, « Pour un composant lexical associatif et stratifié », *D.R.L.A.V.* n° 38, pp. 63-92.

CORBIN Danielle (1990), « Homonymie structurelle et définition des mots construits, vers un dictionnaire dérivationnel *», La définition*, J. Chaurand et F. Mazière éd., Larousse, pp. 175-192.

GARDIN B., LEFEVRE G., MARCELLESI C., MORTUREUX M.-F. (1974), « À propos du sentiment néologique », *Langages* n° 36, pp. 45-52.

GROSS Gaston, 1994, « Classes d'objets et description des verbes », *Langages*, 115, pp. 15-30.

MATHIEU Yvette Yannick., GROSS Gaston, FOUQUERÉ Christophe, « Vers une extraction automatique des néologismes », *Cahiers de lexicologie* n° 72, 1998-1, pp. 199-208

SABLAYROLLES Jean-François, 2000, *La néologie en français contemporain,* examen du concept et analyse de productions néologiques récentes,, coll. Lexica Mots et Dictionnaires, Champion.

SABLAYROLLES Jean-François, 2003, « Le sentiment néologique », *L'innovation lexicale*, J.-F. Sablayrolles éd., Champion, pp. 279-295.

SABLAYROLLES Jean-François « Néologie et dictionnaire(s) comme corpus d'exclusion » (à paraître en 2008), *Néologie et terminologie dans la lexicographie francophone*, J.-F. Sablayrolles éd., coll. Lexica, Champion pp. 19-36.

[21] Cette expérience, présentée dans le numéro 3 de *Neologica* à paraître en 2009, renouvelle une expérience similaire conduite à Limoges en 2000 (Sablayrolles, 2003).